Image Class Prediction by Joint Object, Context, and Background Modeling

Chunjie Zhang, Guibo Zhu, Chao Liang, Yifan Zhang, Qingming Huang, and Qi Tian

***Abstract*— State-of-the-art image classification methods often use spatial pyramid matching or its variants to make use of the spatial layout of visual features. However, objects may appear at various places with different scales and orientations. Besides, traditionally object-centric-based methods only consider objects and the background without fully exploring the context information. To solve these problems, in this paper we propose a novel image classification method by jointly modeling the object, context, and background information (OCB). OCB consists of three components: 1) locate the positions of objects; 2) determine the context areas of objects; and 3) treat the other areas as the background. We use objectness proposal techniques to select candidate bounding boxes. Boxes with high confidence scores are combined to determine objects' positions. To select the context areas, we use candidate boxes that have relatively lower confidence scores compared with boxes for object location selection. The other areas are viewed as the background. We jointly combine the object, context, and background for image representation and classification. Experiments on six data sets well demonstrate the superiority of the proposed OCB method over other spatial partition methods.**

***Index Terms*— Background modeling, context modeling, image class prediction, object modeling.**

I. Introduction

IMAGE classification tries to classify an image based on the objects. The bag-of-visual-words (BoW) model [1] has been widely used for image classification. However, the BoW model does not consider the spatial locations of local features, which limits the discriminative power of the BoW model.

To make use of spatial information, spatial pyramid matching (SPM) [2] is used to divide an image into several regions with a predefined partition strategy. Usually, SPM with three pyramids ($2^L \times 2^L$, $L = 0, 1, 2$) is used. A more flexible SPM strategy (e.g., 3×1) [3] is used when images have large variances. Many works [4]–[15] have greatly improved image classification performances using the SPM technique or its variances. However, these methods simply divide images by hard partitions without considering the locations of objects. The same object may be assigned to different partitions, which degrades the accurate measurement of image similarities.

Object-centric and task-specific methods to cope with the problems of hard partition by detection and segmentation have also been proposed [16]–[20]. These strategies help to represent images more discriminatively for classification. However, more training information (e.g., bounding boxes of objects) is needed. Besides, they are often designed in an iterative way to gradually determine the objects. Moreover, these methods often treat objects and backgrounds separately without considering the context information about the objects and the surrounding areas. The context information should also be considered for image representations.

Contextual correlations [19], [21]–[26] have been widely used to capture more representative information. Some methods try to predefine the context usage strategies. Others try to automatically learn the contextual information from training images. Predefined strategies may not be able to cope with image variations while learning-based methods suffer from heavily computational costs. In recent years, many efficient object detection algorithms [27], [28] have been proposed, which can generate a number of candidate regions. These candidate regions jointly capture the target objects and the surrounding context information. We can use the detected regions for context modeling.

Object-based image representation strategies have proved more efficient than hard partition. However, many object-based methods simply treat the object and background separately without fully exploring the context information. To solve these problems, in this paper we propose a novel image classification method by jointly modeling the object, context, and background information (OCB). We make use of the candidate regions of objectness proposal methods for object and context information extraction. This is achieved by combining the regions with high confidence scores for object selection. The context is determined using the regions with relatively lower confidence scores. The other areas are viewed as the background. In this way, we are able to divide each

Manuscript received May 24, 2016; revised July 30, 2016, August 31, 2016, and September 18, 2016; accepted September 21, 2016. Date of publication September 23, 2016; date of current version February 13, 2018. This work was supported by the National Natural Science Foundation of China under Grant 61303154. This paper was recommended by Associate Editor J. Liang.

C. Zhang is with Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China; with the School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 100049, China; and also with the Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing 100190, China (e-mail: zhangcj@ucas.ac.cn).

G. Zhu is with the Research Center for Brain-inspired Intelligence, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: guibo.zhu@ia.ac.cn).

C. Liang is with the National Engineering Research Center for Multimedia Software, School of Computer, Wuhan University, Wuhan 430072, China (e-mail: cliang@whu.edu.cn).

Y. Zhang is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China (e-mail: yfzhang@nlpr.ia.ac.cn).

Q. Huang is with the School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: qmhuang@ucas.ac.cn).

Q. Tian is with the Department of Computer Sciences, The University of Texas at San Antonio, San Antonio, TX 78249 USA (e-mail: qitian@cs.utsa.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2016.2613125

image into three parts as object, context, and background for joint representation. We evaluate the proposed OCB method with the classification tasks on six image data sets. The experimental results show the effectiveness of the proposed OCB method. Fig. 1 shows the flowchart of the proposed method.

The main contributions of the proposed OCB method lie in three aspects.

1) We jointly model the OCB for image classification. We go beyond object-centric-based strategies by combining the context information about objects. The exploration of context information helps to make use of the discriminative information about objects and the surrounding areas.
2) We use the objectness proposal technique for object and context information selection, which does not need predefined bounding boxes for training. This alleviates the annotation costs and increases the generalization power of the proposed OCB method.
3) OCB can be combined with various object, context, and background representation techniques, e.g., fisher vector (FV) [29], convolutional neural network (CNN) [13], and their variants [30]–[32]. The image representations can then be used for efficient image class prediction.

Both [30] and [32] improve the deep convolutional networks with more efficient pooling strategies, while OCB tries to jointly model the object, context, and background. The contributions of the proposed method are different from [30] and [32]. OCB concentrates on the efficient usage of image parts, while [30] and [32] focus on extracting more discriminative representations from one area. However, OCB can be combined with these two techniques [30], [32] for more efficient object, context, and background representations.

The rest of this paper is organized as follows. The related work is discussed in Section II. In Section III, we give the details of the proposed image classification by jointly modeling the object, context, and background method. We evaluate OCB's performances on six public image data sets in Section IV. Finally, we conclude in Section V.

II. Related Work

The BoW model [1] used the quantized local features to generate histogram representations of images. To consider the spatial information, the SPM technique was used with many variances [2]–[15]. Lazebnik *et al.* [2] proposed to use the $2^L \times 2^L$, $L = 0, 1, 2$ partitions while 3×1 partition was used for the pattern analysis, statistical modeling, and computational learning visual object classes (PASCAL VOC) images [3]. Van Gemert *et al.* [4] softly assigned local features with SPM while Yang *et al.* [5] used sparse coding to encode local features for image representation and classification with improved performances. To make use of the locality information, Wang *et al.* [6] used nearby visual words for sparse reconstruction of local features and reduced the computational cost. Zhang *et al.* [7] implicitly transferred prelearned codebooks for new data sets by adaptively learning the codebook. Gao *et al.* [8] ensured that similar local features should be encoded with smooth parameters with Laplacian sparse coding. Instead of only learning one codebook, Zhang *et al.* [9] learned general and class-specific codebooks to encode the class specific information. To reduce the encoding loss of local features, Zhang *et al.* [10] tried to make use of the residuals to improve the discriminative power of the final image representation. Chatfield *et al.* [11] evaluated the performances of various algorithms and found that the implementation details were very important for efficient classification.

To explore the discriminative information about different color channels, color scale-invariant feature transform (SIFT) was proposed by van de Sande *et al.* [12]. The automatic learning of image representations using deep CNN was also proposed by Krizhevsky *et al.* [13], which greatly improved the classification accuracies. Zhang *et al.* [14] used spatial pyramid coding to consider the spatial locations of features during the codebook learning process. To use the discriminative information for classification, Shih *et al.* [15] tried to learn a collection of part detectors instead.

To avoid dividing the same object into different partitions, the use of detection and segmentation strategies was also explored [16]–[20]. Kim *et al.* [16] tried to divide images based on the specific task. Russakovsky *et al.* [17] proposed the object-centric spatial pooling strategy by iteratively detecting the objects and the classification tasks. Chai *et al.* [18] combined the segmentation technique for image classification with a trilevel cosegmentation scheme. Chen *et al.* [19] combined object detection and classification into a unified framework and improved the detection and classification performances. Angelova and Zhu [20] targeted the fine-grained classification problem by object detection and segmentation.

The context information has been widely used for various applications [19], [21]–[26]. Fei-Fei and Perona [21] tried to learn a hierarchical model for scene classification. Ding *et al.* [22] combined the hierarchical context information for image annotation. Rasiwasia and Vasconcelos [23] tried to map images with semantic representations using holistic context models, while Zheng *et al.* [24] encoded local features with graph constraints. To encode high-order correlations, Liu *et al.* [25] generated a hypergraph for classification. Zhang *et al.* [26] also used deep neural networks to fuse multiple semantic cues to classify different events. However, determining the locations of objects required labeled regions for training, which costed a lot of human labor. We could make use of other objectness detection algorithms [27], [28] to first generate a number of candidate regions and then select the objects and context. Besides, there were also many efficient image representation methods (e.g., FV [29] and CNN [13]), which could be combined to represent images.

It has been proved very useful to jointly use the discriminative representations with effective models [30]–[37]. Gong *et al.* [30] applied orderless pooling over convolutional networks with multiscales for image representation. Simonyan and Zisserman [31] trained very deep convolutional networks for large-scale recognition while He *et al.* [32] combined spatial pyramid pooling in deep convolutional networks. Nilsback and Zisserman [33] generated the flower-specific codebook for classification, while Varma and Ray [35] tried to balance the discriminative power. Yuan and Yan [36]

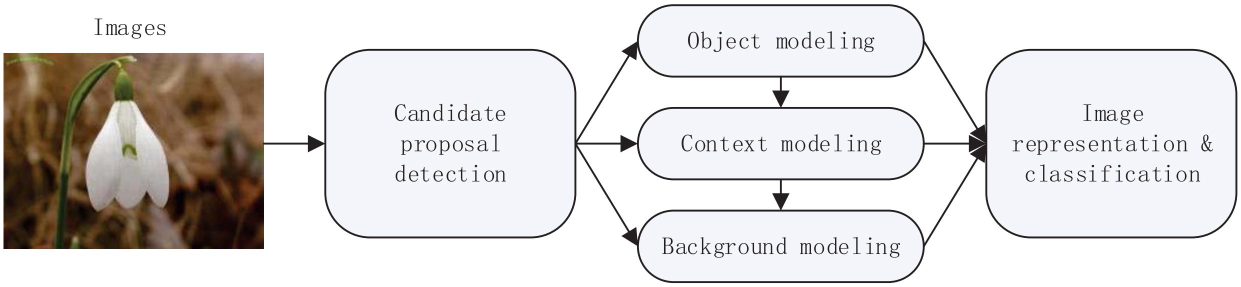

Images
Candidate proposal detection
Object modeling
Context modeling
Background modeling
Image representation & classification

Fig. 1. Flowchart of the proposed joint object, context, and background modeling method for image classification.

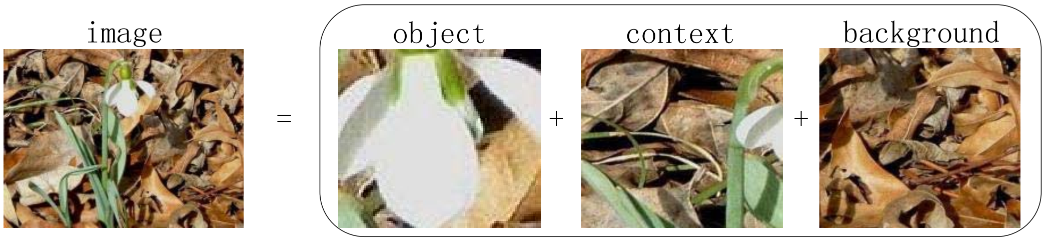


Fig. 2. Toy partition of the object, context, and background of a flower image.

targeted the multitask classification problem by joint sparse representation. Oquab *et al.* [37] used CNN to transfer midlevel image representations with good performance.

To improve the classification performance, researchers have made use of various information [11], [20], [38]–[43]. Hu *et al.* [38] measured the histogram similarities of images with bin ratio-based methods and improved the performance compared with Euclidean distance. Angelova and Zhu [20] combined object detection and segmentation for fine-grained recognition. A generative model was proposed by Fei-Fei *et al.* [42], while Nilsback and Zisserman [43] classified flower images with a large number of classes. Chatfield *et al.* [11] evaluated different classification methods and found that the implementation details played important roles for efficient recognition.

The direct usage of local features was also explored [44]–[46]. Boiman *et al.* [44] measured image similarities using local features with heavy computational cost. Athitsos *et al.* [45] used a cascade of the approximate similarity measurement method to speed up the computation of similarities. McCann and Lowe [46] proposed to use local naïve Bayes nearest neighbor search instead. To improve the classification performances, many well-designed algorithms [30], [40], [47]–[59] were proposed and tested on various data sets.

III. Joint Modeling of Object, Context, and Background for Image Classification

In this section, we give the details of the proposed image classification method by joint object, context, and background modeling. We first determine the objectness proposals and then use them for image representation. The proposals with high confidence scores are used for object modeling, while proposals with relatively lower confidence scores are combined for context modeling. The other image areas are treated as background. We then concatenate the OCB for image representation and train classifiers to predict the classes of images. Fig. 2 shows a toy partition of the object, context, and background of a flower image.

A. Object Modeling

To model an object, we first determine its location. However, it is often very hard to accurately select the object's location. Fortunately, a number of objectness proposal methods [27], [28] have been proposed, which can generate a number of proposals to cover the objects in a given image. We can harvest this information for object modeling.

We combine the objectness proposals generated by binarized normed gradients (BING) [27] and EdgeBoxes [28] together for image representation. The confidence scores of objectness proposals are first normalized for the BING- and EdgeBoxes-based methods, respectively, and then combined together. Formally, let $\boldsymbol{S} = [s_1, s_2, \ldots, s_N]$ be the sorted set of confidence scores with s_1 as the highest score and s_N as the lowest score. For each image, let $\boldsymbol{X} = [\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_N}]$ be the corresponding set of objectness proposals where N is the proposal number, $\boldsymbol{x_n}, n = 1, \ldots, N$ is the nth bounding box. The confidence score indicates the probability that one area contains the object. If the confidence score is high, the bounding box probably covers a large part of the object. However, only using the bounding box with the largest confidence score is not enough. It is more effective to combine several bounding boxes for joint modeling.

We propose to use the first M bounding boxes with relatively higher confidence scores. We use the selected bounding boxes with high confidence scores to generate an $M \times M$ similarity matrix $\boldsymbol{W}$. We use the intersection over union (IoU) scores [27], [28] to measure the similarities of different proposals. IoU is defined as

$$W_{i,j} = \frac{|\boldsymbol{x_i} \bigcap \boldsymbol{x_j}|}{|\boldsymbol{x_i} \bigcup \boldsymbol{x_j}|}. \tag{1}$$

We cluster the M bounding boxes into $K(K < M)$ clusters using K-means clustering. The input for K-means clustering is the location information about bounding boxes with the similarity between two bounding boxes defined as (1). By clustering the bounding boxes, we can ensure that the selected bounding boxes within one cluster are spatially nearby and may probably concentrate on the same object. For the kth cluster, we use the corresponding bounding boxes to generate the location of the object as

$$\boldsymbol{A}_{\text{obj}} = \boldsymbol{x}_{k,1} \bigcap \boldsymbol{x}_{k,2} \bigcap \cdots \bigcap \boldsymbol{x}_{k,k_m} \tag{2}$$

where $\boldsymbol{x}_{k,j}, j = 1, \ldots, k_m$ is the jth bounding box of the kth cluster. Since we only use the bounding boxes with high confidence scores for object location selection, we are probably able to choose the area that belongs to the corresponding object. Note that we do not try to select the exact locations of objects but try to determine the areas that lie on the objects. Besides, the clustering scheme also ensures that the select object area is not too small. Moreover, unlike detection-based methods that select rectangular areas, the proposed object selection method combines a number of rectangular areas. This helps to reduce the noisy information and represent the objects better without segmentation. After determining the locations of objects, we can use them to select the surrounding context and background for image representation.

B. Context Modeling

The context information has been shown to be very important for efficient classification. Instead of only using the bounding boxes with relatively higher confidence scores, we also use the boxes with relatively lower scores for context modeling. This is different from the object modeling process where only boxes with high scores are selected. The object detection task tries to determine the exact locations of objects. However, only selecting the objects is not enough for reliable classification, especially when the objects are visually similar (e.g., different classes of flowers with similar appearances); the context information should also be incorporated. Context information refers to the discriminative areas that contain both the objects and the surrounding information. This context information can help to separate different objects and probably lies in the boxes with relatively lower scores. Other object detection methods simply discard the low scored proposals. However, we believe this information can be used for efficient representation.

We use the top $\widetilde{M}$ bounding boxes for context modeling. In this way, we can obtain an image region by combining these boxes as

$$\widetilde{\boldsymbol{A}} = \boldsymbol{x}_1 \bigcup \boldsymbol{x}_2 \bigcup \cdots \bigcup \boldsymbol{x}_{\widetilde{M}} \tag{3}$$

with $\widetilde{M} \geq M$. $\widetilde{\boldsymbol{A}}$ captures useful information about both objects and the surrounding areas. If we set $\widetilde{M} = M$, the proposed method would degenerate to the traditional object centric-based methods. To select the context area, we exclude the object area from $\widetilde{\boldsymbol{A}}$ as

$$\boldsymbol{A}_{\text{context}} = \widetilde{\boldsymbol{A}} - \boldsymbol{A}_{\text{obj}}. \tag{4}$$

Note that the selected context area $\boldsymbol{A}_{\text{context}}$ cannot be directly computed by $\boldsymbol{x}_{M+1} \bigcup \cdots \bigcup \boldsymbol{x}_{\widetilde{M}}$. Let $\widehat{\boldsymbol{A}} = \boldsymbol{x}_{M+1} \bigcup \cdots \bigcup \boldsymbol{x}_{\widetilde{M}}$. If we use $\widehat{\boldsymbol{A}}$ as the context information, we would leave some areas unconsidered. Since these areas probably cover parts of the objects and surrounding areas, the discriminative power of the selected areas would be severely degraded.

Note that we use different operations to combine the information about objectness proposals for object and context modeling. When modeling the objects, we need to make sure the selected regions only contain the objects and remove the noisy information as much as possible. Hence, the $\bigcap$ operation (intersection operation) is used. However, when modeling context information, we need to combine the surrounding information as much as possible. Hence, the $\bigcup$ operation (union operation) is used. To avoid the selected context area being too large, we only use boxes with relatively lower scores instead of all the boxes. This helps to reduce the contamination of noisy information during the context modeling process.

C. Background Modeling

We can model the background after the object and context are selected. The background region is defined as the other image areas that do not belong to either the objects or the context. Formally, if $\boldsymbol{I}$ is the whole image region, the background can be determined as

$$\begin{aligned}\boldsymbol{A}_{\text{background}} &= \boldsymbol{I} - \widetilde{\boldsymbol{A}} \\ &= \boldsymbol{I} - \boldsymbol{A}_{\text{obj}} - \boldsymbol{A}_{\text{context}}.\end{aligned} \tag{5}$$

We can use various representation methods (e.g., sparse coding, FV, and CNN) for object, context, and background modeling. The histogram representations of object, context, and background are concatenated to get the image representation $\boldsymbol{h}$, which can then be used for classifier training and image class prediction.

D. Image Class Prediction

Let $(\boldsymbol{h}_p, y_p), p = 1, \ldots, P$ be the set of training images with P as the number of images. $\boldsymbol{h}_p$ is the concatenated image representation for the pth image. We train linear classifiers to predict the image class as

$$\widetilde{y_p} = \boldsymbol{w}^T \boldsymbol{h}_p \tag{6}$$

by minimizing the quadratic hinge loss with L_2 constraint as

$$\boldsymbol{w} = argmin_{\boldsymbol{w}} \sum_{p=1}^{P} (\max(0, 1 - y_p \boldsymbol{w}^T \boldsymbol{h}_p))^2 + \lambda \|\boldsymbol{w}\|_2^2 \tag{7}$$

where λ is the parameter for controlling the influence of the regularization term. Quadratic hinge loss is used for smooth reasons. After the classifiers are learned, we predict the image class with the category that has the largest response. Other classifier training methods can also be used for image class prediction. Algorithm 1 gives the procedures of the proposed image classification by the joint object, context, and background modeling method.

IV. Experiments

To evaluate the effectiveness of the proposed OCB method, we conduct classification experiments on the Flower-17 [33], Caltech-256 [34], PASCAL VOC 2007 [3], Flower-102 [43], MIT-Indoor [47], and Caltech-101 data sets [42]. The codebook size is set to 1000 for local feature-based methods.

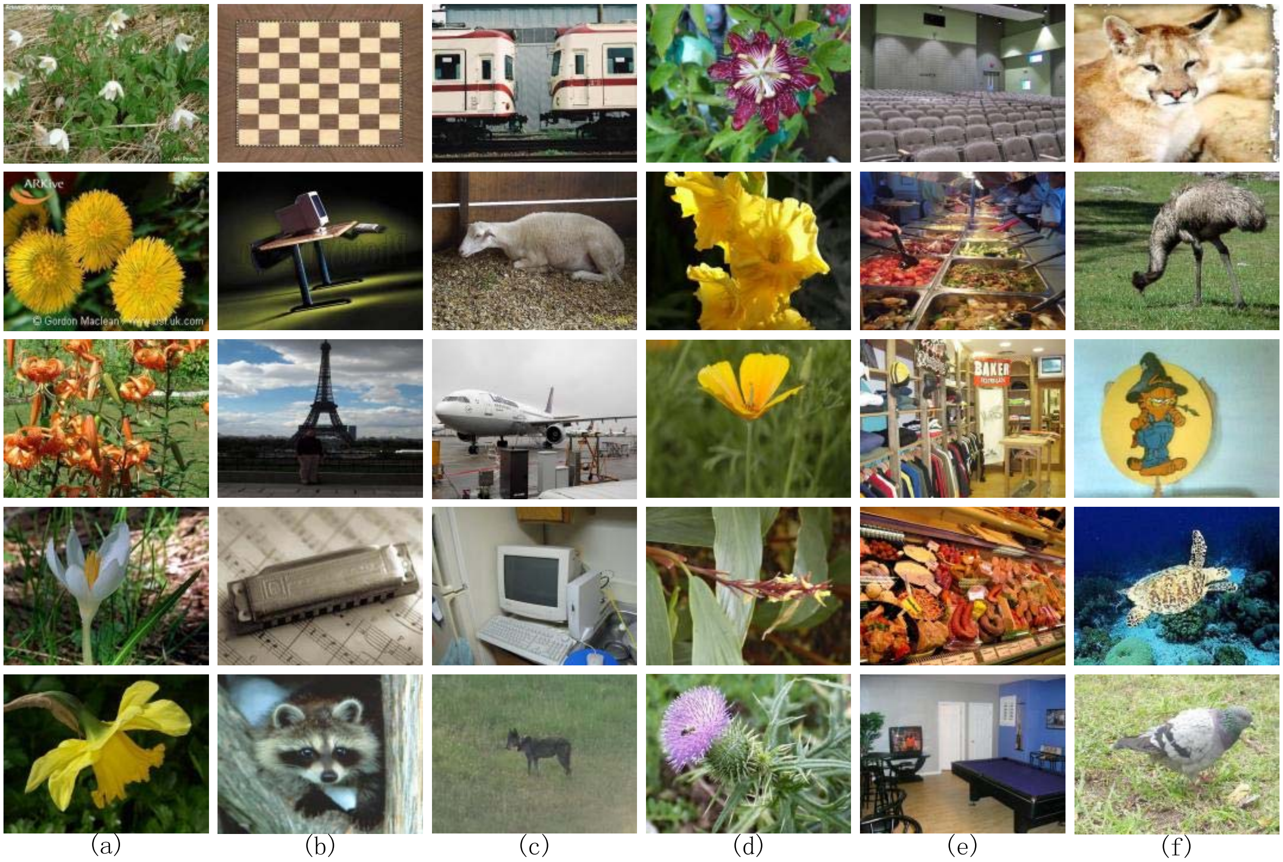


Fig. 3. Example images of (a) Flower-17, (b) Caltech-256 (45 images), (c) PASCAL VOC 2007, (d) Flower-102, (e) MIT-Indoor, and (f) Caltech-101 data sets.

```
Algorithm 1 Procedures of the Proposed Image Classification by Joint Object, Context, and Background Modeling Method
Input:
    Training images, the objectness proposals $\boldsymbol{X}$ with the corresponding confidence scores $S$, the parameters $\widetilde{M}$ and M.
Output:
    The learned classifiers;
    Select the object areas using Eq. 2.
    Determine the context areas with Eq. 4.
    Calculate the background areas with Eq. 5.
    Learn the classifiers for image class prediction by optimizing over Eq. 7.
    return The learned classifiers.
```

We use both the local feature and CNN-based methods on the Caltech-101, MIT-Indoor, and PASCAL VOC 2007 data sets. We use the image completion technique [48] to handle the nonrectangular problem. CNN-S model [11] is used to extract the corresponding CNN features. Fig. 3 shows some example images of the six data sets.

A. Flower-17 Data Set

This data set has 17 classes of flowers (*buttercup, colts foot, daffodil, daisy, dandelion, fritillary, iris, pansy, sunflower, windflower, snowdrop, lily valley, bluebell, crocus, tigerlily, tulip*, and *cowslip*) with different shapes and colors. There are 1360 images with 80 images for each class. For local feature-based image representation, we densely extract color SIFT features [12] with multiscales and overlap. We set the minimum scale to 16×16 pixels and the overlap to 8 pixels. We follow the experimental setup as in [33] and use 40/20/20 images, respectively, for train/validate/test. The classification rate is used for performance comparison by directly comparing with the results reported by other methods. Table I gives the classification performances on the Flower-17 data set. We give the performances of OCB using sparse coding (OCB-SC) and FV (OCB-FV) for local feature encoding, respectively. To show the advantages of adding context information, we also give the performances of only using object and background using sparse coding (OB-SC) and FV (OB-FV).

TABLE I

Classification Rate Comparison on the Flower-17 Data Set

Algorithm	Performance
ICT [7]	91.37± 0.72
LR-GCC [9]	91.52± 1.24
Nilsback [33]	71.76± 1.76
Varma [35]	82.55± 0.34
KMTJSRC-CG [36]	88.90± 2.30
mTDP [51]	94.89± 0.90
OB-SC	89.57± 0.73
OB-FV	92.41± 0.82
OCB-SC	94.50± 0.86
OCB-FV	97.84± 0.78

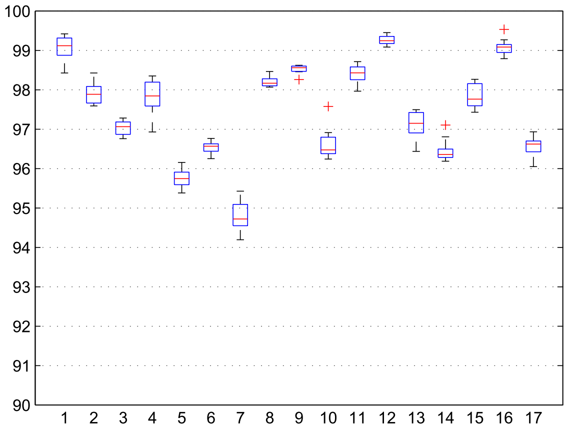


Fig. 4. Boxplot of per-class performances of OCB-FV on the Flower-17 data set. The numbers from 1 to 17 in the horizontal row represent buttercup, colts foot, daffodil, daisy, dandelion, fritillary, iris, pansy, sunflower, windflower, snowdrop, lily valley, bluebell, crocus, tigerlily, tulip, and cowslip, respectively.

From Table I, we can draw four conclusions. First, the proposed OCB method is able to improve the classification accuracy over many baseline methods. Second, OCB is able to represent images more discriminatively than the SPM strategies [7], [33], [35], [36] by jointly modeling the OCB. Different flower images may have visually similar features. Hence, the efficient exploration of context information is very important. Third, by encoding local features more finely, OCB-FV outperforms OCB-SC by 3%. FV can capture higher order information than sparse coding and improve the classification accuracy, as shown in Fig. 4. Fourth, OCB-SC performs comparable to [49], which uses task-specific pooling scheme along with CNN-based features for image representation. However, OCB outperforms [49] when FV is used for representation. We also show the boxplot of the per-class performances of OCB-FV in Fig. 4. We can see from Fig. 4 that the proposed method is able to obtain stable performances on different classes.

B. Caltech-256 Data Set

This data set has 29 780 images of 256 classes. We densely extract SIFT features with multiscales (the minimum scale is 16×16 pixels) and overlap (6 pixels). We randomly select 15, 30, 45, and 60 images per class for training and use the remaining images for testing. We repeat the random selection process ten times. Table II gives the performance comparisons of OCB with other methods [4]–[6], [8], [9], [29], [34] on the Caltech-256 data set.

Compared with SPM, which simply divides images with predefined rules, OCB can adaptively represent the OCB more efficiently. This helps OCB to improve over SPM-based methods by about 6%. Besides, OCB-SC-based method can also outperform FV [29], which considers both the first-order and second-order information during the encoding process. The performance can be further improved when OCB is combined with FV. Moreover, the proposed method with a single codebook also improves over low-rank sparse coding with general and class-specific codebooks [9], which learns a number of codebooks. Finally, OCB also improves the classification performance over naïve Bayes nearest neighbor, which uses local features directly. The improvement also increases with the number of training images as we can make use of the more efficient OCB-based representations for classifier training. The classifiers are more discriminative when trained with more images. The results on the Caltech-256 data set again prove the effectiveness of the proposed joint object, context, and background modeling method.

C. PASCAL VOC 2007 Data Set

This data set has 20 classes (*aeroplane, bicycle, boat, bottle, bus, bird, car, cat, cow, chair, dining table, dog, horse, person, sheep, motorbike, train, potted plant, soft*, and *tv/monitor*) of more than 10 000 images. The images are provided with train/validate/test splits. We follow the setup given in [3], [6], [11], [17], [29], and [37] and train the classifiers using the train split and select the optimal parameters with the validate split. The train and validate splits are then merged together to retrain the classifier with the optimal parameters. The final performance is evaluated by applying the learned classifiers on the test split. Average precision is used for performance evaluation.

Table III gives the average precision per class and mean average precision comparisons of OCB with other methods [3], [6], [11], [17], [29], [37] on the PASCAL VOC 2007 data set. We can see that the proposed OCB is able to achieve superior performances compared with many baseline methods. Especially, OCB improves over SPM again. The improvement mainly lies in the animal classes (e.g., cat, cow, and dog). The relative places of animals are more variable than rigid objects. Hence, explicit modeling of objects can help represent images discriminatively. Besides, OCB also improves over OCP, which treats images using object and background only without considering the context information. Since images of the PASCAL VOC 2007 data set are more difficult to classify than the other two data sets, due to variations and occlusions, the modeling of context information is very important for the final classification. Moreover, OCB is able to consistently improve the accuracy when combined with FV, because the local feature can be encoded more efficiently. Finally, we can further improve the classification performance using the discriminative CNN-based representations.

D. Flower-102 Data Set

This data set has 102 classes of flower images with a total number of 8189 images, which is an extended data set of the Flower-17 data set. We follow the experimental setup as in [43] and use 10/10/remaining images for train/validate/test, respectively. Color SIFT features are used for local region representations. We give the classification rate comparisons of OCB with other baseline methods in Table IV.

Compared with hard partition [43], the proposed OCB method is able to capture the object information for better classification. This helps to improve the performance over SPM by about 10%. OCB-SC also outperforms [20], which uses detection and segmentation techniques for classification. Only modeling the objects is not enough for reliable classification, the context information should also be incorporated.

TABLE II
PERFORMANCE COMPARISONS ON THE CALTECH-256 DATA SET

Methods	15 images	30 images	45 images	60 images
KSPM [34]	–	34.10	–	–
KSPM [5]	23.34 ± 0.42	29.51 ± 0.52	–	–
ScSPM [5]	27.73 ± 0.51	34.02 ± 0.35	37.46 ± 0.55	40.14 ± 0.91
LR-GCC [9]	39.21 ± 0.48	45.87 ± 0.41	–	–
KC [4]	–	27.17 ± 0.46	–	–
LLC [6]	34.36	41.19	45.31	37.79 ± 0.42
LScSPM[8]	30.00 ± 0.14	35.74 ± 0.10	38.54 ± 0.36	40.32 ± 0.32
FV[29]	38.5	47.4	52.1	54.80 ± 0.40
NBNN [46]	35.20	42.80	–	–
OCB-SC	40.56 ± 0.53	49.87 ± 0.50	54.18 ± 0.49	57.60 ± 0.51
OCB-FV	44.03 ± 0.46	53.15 ± 0.44	57.84 ± 0.40	59.03 ± 0.45

TABLE III
PERFORMANCE COMPARISONS ON THE PASCAL VOC 07 DATA SET

object class	LLC [6]	Best07[3]	FV [29]	DECAF[11]	CNN[11]	PRE[37]	SPM[17]	OCP[17]	OCB-SC	OCB-FV	OCB-CNN
airplane	74.8	77.5	80.0	87.4	95.3	88.5	72.5	74.2	76.9	82.1	96.2
bicycle	65.2	63.6	67.4	79.3	90.4	81.5	56.3	63.1	67.1	69.3	91.7
bird	50.7	56.1	51.9	84.1	92.5	87.9	49.5	45.1	53.6	54.4	93.3
boat	70.9	71.9	70.9	78.4	89.6	82.0	63.5	65.9	72.1	72.6	90.1
bottle	28.7	33.1	30.8	42.3	54.4	47.5	22.4	29.5	32.8	34.2	56.8
bus	68.8	60.6	72.2	73.7	81.9	75.5	60.1	64.7	70.6	73.8	82.3
car	78.5	78.0	79.9	83.7	91.5	90.1	76.4	79.2	81.2	81.7	91.9
cat	61.7	58.8	61.4	83.7	91.9	87.2	57.5	61.4	64.4	64.5	92.4
chair	54.3	53.5	56.0	54.3	64.1	61.6	51.9	51.0	55.9	58.3	66.2
cow	48.6	42.6	49.6	61.9	76.3	75.7	42.2	45.0	52.6	53.6	78.5
table	51.8	54.9	58.4	70.2	74.9	67.3	48.9	54.8	54.7	60.7	76.3
dog	44.1	45.8	44.8	79.5	89.7	85.5	38.1	45.4	48.5	47.2	91.1
horse	76.6	77.5	78.8	85.3	92.2	83.5	75.1	76.3	79.2	81.4	94.7
motorbike	66.9	64.0	70.8	77.2	86.9	80.0	62.8	67.1	68.3	72.3	89.2
person	83.5	85.9	85.0	90.5	95.2	95.6	82.9	84.4	85.8	86.9	95.6
plant	30.8	36.3	31.7	51.1	60.7	60.8	20.5	21.8	35.2	35.4	62.4
sheep	44.6	44.7	51.0	73.8	82.9	76.8	38.1	44.3	46.6	53.8	84.1
sofa	53.4	50.9	56.4	57.0	68.0	58.0	46.0	48.8	55.8	58.7	70.4
train	78.2	79.2	80.2	86.4	95.5	90.4	71.7	70.7	81.9	82.6	96.2
tv	53.5	53.2	57.5	68.0	74.4	77.9	50.5	51.7	56.8	61.1	77.8
mAP	59.3	59.4	61.7	73.4	82.4	77.7	54.3	57.2	62.0	64.2	83.9

TABLE IV
MEAN CLASSIFICATION RATE COMPARISONS ON THE FLOWER-102 DATA SET

Methods	Classification rate
LR-GCC [9]	75.7
Nilsback [44]	72.8
Hu [38]	86.8
KMTJSRC-CG [36]	74.1
Det+Seg [39]	80.7
TriCoS [18]	85.2
OCB-SC	84.6
OCB-FV	91.3

Besides, using the FV for local feature encoding can consistently improve the performance over sparse coding. Moreover, discriminatively trained classifiers also help to get higher classification accuracy compared with the reconstructions-based method [36].

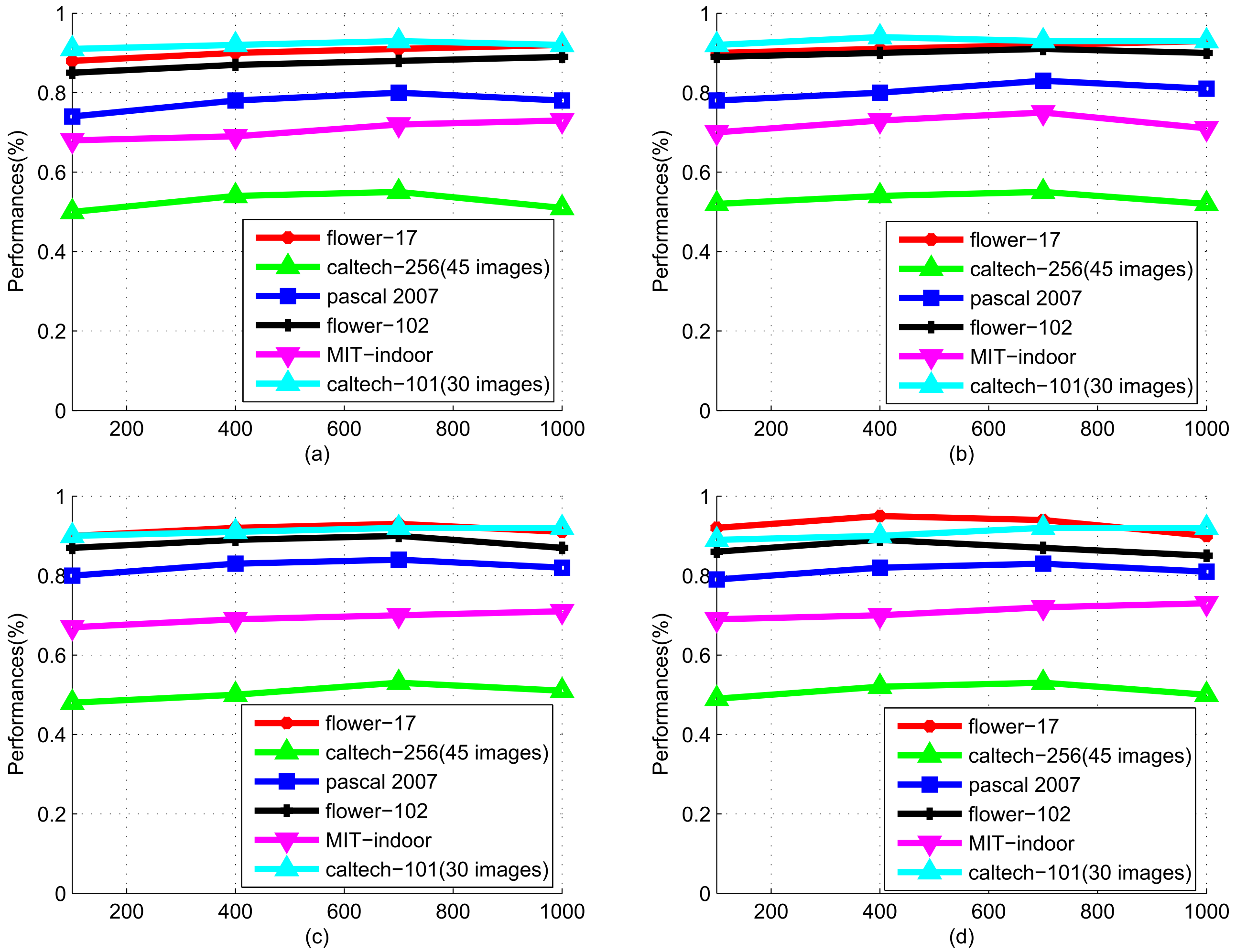
E. MIT-Indoor Data Set

TABLE V
MEAN CLASSIFICATION RATE COMPARISONS ON THE MIT-INDOOR DATA SET

Methods	Classification rate
KSPM [2]	34.40
Quattoni [49]	26.50
mTDP [51]	75.61
Doersch [52]	66.87
Gong [53]	68.90
Lin [54]	68.50
Razavian [55]	69.00
Zhou [56]	70.08
CENTRIST [57]	36.90
LPR-LIN [58]	44.84
OCB-SC	52.60
OCB-CNN	75.84

The MIT-Indoor data set has 67 classes of indoor images. There are 15620 images in total. This data set is hard to separate as the inter-class variances are small compared with other data sets. Hence, how to model the images more efficiently becomes a question that needed to be solved. Since Quattoni and Torralba [47] have provided the data split strategy, we follow this setup and use 80 images per-class for training. Table V gives the performance comparisons on this data set.

Compared with SPM-based methods [2], [47], [55], OCB is more flexible for representation, which helps to model images more appropriately. Besides, OCB with local features cannot



(a)

(b)

(c)

(d)

Fig. 5. Influences of the proposal numbers (M, $\widetilde{M}$) for object and context modeling on the Flower-17, Caltech-256 (45 images), PASCAL VOC 2007, Flower-102, MIT-Indoor, and Caltech-101 data sets (30 images) with (a) $\widetilde{M} = 1000$, (b) $\widetilde{M} = 1500$, (c) $\widetilde{M} = 2000$, and (d) $\widetilde{M} = 2500$. The horizontal row indicates M.

perform as deep learning-based methods [30], [50]–[53]. The deeply learned representations are more discriminative. However, when combined with deep learning-based features, OCB-CNN is able to outperform these methods. Since images of the indoor scenes are cluttered and hard to classify, it is relatively more difficult to model the object and context along with the background. Hence, the relative improvement over mTDP on the MIT Indoor data set is not as large as that on the other data sets.

F. Caltech-101 Data Set

The Caltech-101 data set has 101 classes of objects and 1 noisy class of 9144 images. The number of images per class ranges from 31 to 800. We randomly select 15 and 30 training images for training and use the other images for testing. The random selection process is repeated ten times. We give the performance comparisons of OCB with the baseline methods in Table VI. From Table VI, we can see that the proposed OCB method again improves the classification accuracy over many baseline methods. With the development of CNN-based features, the image classification performance can be greatly improved. Besides, by representing images more properly, the similarities between images can be more accurately measured. Moreover, training discriminative classifiers also helps to boost the performance over image-to-class similarities with local features [44]. Finally, we are able to improve over task-specific representations [47]. The results again prove the usefulness of the proposed OCB method for image classification.

G. Influences of Proposal Numbers

We show the influences of the proposal numbers for object and context modeling in Fig. 5. If M is set to a large number, the resulting object area would also be large, and vice versa. Similarly, If $\widetilde{M}$ is relatively large, the object and context areas would cover the whole image. Hence, choosing the proper proposal numbers M and $\widetilde{M}$ is very important for the final performances. Besides, different images may require different proposal numbers. For example, for the flower images, since the flowers mainly appear on the central areas of images that are relatively easier to detect, relatively smaller proposal numbers would be enough to obtain satisfactory classification performances. However, for the images of the PASCAL VOC 2007 data set, we need more candidate areas to locate the objects and the context information. Hence, we set (M, $\widetilde{M}$)

TABLE VI
MEAN CLASSIFICATION RATE COMPARISONS ON THE CALTECH-101 DATA SET

Methods	15 training	30 training
KSPM [2]	56.40	64.40 ± 0.80
ScSPM [5]	67.00 ± 0.45	73.20 ± 0.54
KC [4]	–	64.14 ± 1.18
VGG [31]	–	92.70 ± 0.50
NBNN [46]	65.00 ± 1.14	70.40
mTDP [51]	–	93.68 ± 0.50
LLC [6]	65.43	73.44
KMTJSRC-CG [36]	65.00 ± 0.70	–
SBLM [59]	71.68 ± 1.12	77.50 ± 0.77
SVM-KNN [60]	59.10 ± 0.60	66.20 ± 0.50
CNN [61]	–	84.77 ± 0.70
OCB-SC	73.32 ± 0.96	81.68 ± 0.83
OCB-CNN	84.51 ± 0.84	95.05 ± 0.92

to (300, 1000) for the five data sets and K to 20 except the PASCAL VOC 2007 data set, which is set to (500, 1200) and 30, respectively. Moreover, we can see from Fig. 5 that the performances are relatively stable. This again proves the usefulness of the proposed OCB method.

V. CONCLUSION

In this paper, we proposed an efficient image classification method by modeling the object, context, and background jointly. We made use of the objectness proposal technique using the candidate bounding boxes for object and context selection. The object, context, and background representations were jointly combined for image representation. In this way, we could measure the similarities of images more effectively and accurately. Besides, the modeling of contextual information of objects also made the final representation robust to image variations. We evaluated the effectiveness of OCB for image classification on six data sets. The experimental results demonstrated the superiorities of OCB over other spatial partition methods.

REFERENCES

[1] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. Int. Conf. Comput. Vis.*, 2003, pp. 1470–1477.

[2] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. Comput. Vis. Pattern Recognit.*, New York, NY, USA, 2006, pp. 2169–2178.

[3] M. Everingham, A. Zisserman, C. Williams, and L. Van Gool, "The PASCAL visual object classes challenge 2007 (VOC 2007) results," Pascal Challenge, London, U.K., Tech. Rep. 1, 2007.

[4] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek, "Visual word ambiguity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1271–1283, Jul. 2010.

[5] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, 2009, pp. 1794–1801.

[6] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. Comput. Vis. Pattern Recognit.*, 2010, pp. 3360–3367.

[7] C. Zhang, J. Cheng, J. Liu, J. Pang, Q. Huang, and Q. Tian, "Beyond explicit codebook generation: Visual representation using implicitly transferred codebooks," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5777–5788, Dec. 2015.

[8] S. Gao, I. W.-H. Tsang, and L.-T. Chia, "Laplacian sparse coding, hypergraph Laplacian sparse coding, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 92–104, Jan. 2013.

[9] C. Zhang, C. Liang, L. Li, J. Liu, Q. Huang, and Q. Tian, "Fine-grained image classification via low-rank sparse coding with general and class-specific codebooks," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published. [Online]. Available: http://ieeexplore.ieee.org/document/7448920/

[10] C. Zhang, J. Liu, C. Liang, Q. Huang, and Q. Tian, "Image classification using Harr-like transformation of local features with coding residuals," *Signal Process.*, vol. 93, no. 8, pp. 2111–2118, Aug. 2013.

[11] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. BMVC*, 2014, pp. 1–12.

[12] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1582–1596, Sep. 2010.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[14] C. Zhang, S. Wang, Q. Huang, J. Liu, C. Liang, and Q. Tian, "Image classification using spatial pyramid robust sparse coding," *Pattern Recognit. Lett.*, vol. 34, no. 9, pp. 1046–1052, 2013.

[15] K. J. Shih, I. Endres, and D. Hoiem, "Learning discriminative collections of part detectors for object recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 8, pp. 1571–1584, Aug. 2015.

[16] S. Kim, S. Nowozin, P. Kohli, and C. D. Yoo, "Task-specific image partitioning," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 488–500, Feb. 2013.

[17] O. Russakovsky, Y. Lin, K. Yu, and L. Fei-Fei, "Object-centric spatial pooling for image classification," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 1–15.

[18] Y. Chai, E. Rahtu, V. S. Lempitsky, L. Van Gool, and A. Zisserman, "TriCoS: A tri-level class-discriminative co-segmentation method for image classification," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 794–807.

[19] Q. Chen, Z. Song, J. Dong, Z. Huang, Y. Hua, and S. Yan, "Contextualizing object detection and classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 1, pp. 13–27, Jan. 2015.

[20] A. Angelova and S. Zhu, "Efficient object detection and segmentation for fine-grained recognition," in *Proc. Comput. Vis. Pattern Recognit.*, 2013, pp. 811–818.

[21] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. Comput. Vis. Pattern Recognit.*, 2005, pp. 524–531.

[22] X. Ding, B. Li, W. Xiong, W. Gao, W. Hu, and B. Wang, "Multi-instance multi-label learning combining hierarchical context and its application to image annotation," *IEEE Trans. Multimedia*, vol. 18, no. 8, pp. 1616–1627, Aug. 2016.

[23] N. Rasiwasia and N. Vasconcelos, "Holistic context models for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 902–917, May 2012.

[24] M. Zheng *et al.*, "Graph regularized sparse coding for image representation," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1327–1336, May 2011.

[25] Q. Liu, Y. Sun, C. Wang, T. Liu, and D. Tao, "Elastic net hypergraph learning for image clustering and semi-supervised classification," *CoRR*, arXiv:1603.01096, Mar. 2016.

[26] X. Zhang *et al.*, "Deep fusion of multiple semantic cues for complex event recognition," *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1033–1046, Mar. 2016.

[27] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "BING: Binarized normed gradients for objectness estimation at 300 fps," in *Proc. Comput. Vis. Pattern Recognit.*, 2014, pp. 3286–3293.

[28] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 391–405.

[29] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *Int. J. Comput. Vis.*, vol. 105, no. 3, pp. 222–245, Dec. 2013.

[30] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 392–407.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, arXiv:1409.1556 [cs.CV], Apr. 2015.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[33] M.-E. Nilsback and A. Zisserman, "A visual vocabulary for flower classification," in *Proc. Comput. Vis. Pattern Recognit.*, 2006, pp. 1447–1454.

[34] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," Division Chemistry Chemical Eng., California Instit. Technol., CalTech, Pasadena, CA, USA, Tech. Rep. 1, 2007.

[35] M. Varma and D. Ray, "Learning the discriminative power-invariance trade-off," in *Proc. Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[36] X.-T. Yuan and S. Yan, "Visual classification with multi-task joint sparse representation," in *Proc. Comput. Vis. Pattern Recognit.*, 2010, pp. 3493–3500.

[37] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. Comput. Vis. Pattern Recognit.*, 2014, pp. 1717–1724.

[38] W. Hu *et al.*, "Bin ratio-based histogram distances and their application to image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 12, pp. 2338–2352, Dec. 2014.

[39] C. Li *et al.*, "Spatially regularized streaming sensor selection," in *Proc. AAAI*, 2016, pp. 3871–3879.

[40] C. Li, Q. Liu, W. Dong, X. Zhu, J. Liu, and H. Lu, "Human age estimation based on locality and ordinal information," *IEEE Trans. Cybern.*, vol. 45, no. 11, pp. 2522–2534, Nov. 2015.

[41] C. Li, Q. Liu, J. Liu, and H. Lu, "Learning ordinal discriminative features for age estimation," in *Proc. Comput. Vis. Pattern Recognit.*, 2012, pp. 2570–2577.

[42] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," in *Proc. Comput. Vis. Pattern Recognit. Workshop*, 2004, p. 178.

[43] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. ICCVGIP*, 2008, pp. 722–729.

[44] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *Proc. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[45] V. Athitsos, J. Alon, and S. Sclaroff, "Efficient nearest neighbor classification using a cascade of approximate similarity measures," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 486–493.

[46] S. McCann and D. G. Lowe, "Local naive Bayes nearest neighbor for image classification," in *Proc. Comput. Vis. Pattern Recognit.*, 2012, pp. 3650–3656.

[47] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *Proc. Comput. Vis. Pattern Recognit.*, 2009, pp. 413–420.

[48] I. Drori, D. Cohen-Or, and H. Yeshurun, "Fragment-based image completion," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 303–312, 2003.

[49] G.-S. Xie, X.-Y. Zhang, X. Shu, S. Yan, and C.-L. Liu, "Task-driven feature pooling for image classification," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 1179–1187.

[50] C. Doersch, A. Gupta, and A. A. Efros, "Mid-level visual element discovery as discriminative mode seeking," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 494–502.

[51] D. Lin, C. Lu, R. Liao, and J. Jia, "Learning important spatial pooling regions for scene classification," in *Proc. Comput. Vis. Pattern Recognit.*, 2014, pp. 3726–3733.

[52] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. Comput. Vis. Pattern Recognit.*, 2014, pp. 806–813.

[53] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 487–495.

[54] J. Wu and J. M. Rehg, "Centrist: A visual descriptor for scene categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1489–1501, Aug. 2011.

[55] F. Sadeghi and M. F. Tappen, "Latent pyramidal regions for recognizing scenes," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 228–241.

[56] C. Zhang, J. Liu, Q. Tian, C. Xu, H. Lu, and S. Ma, "Image classification by non-negative sparse coding, low-rank and sparse decomposition," in *Proc. CVPR*, 2011, pp. 1673–1680.

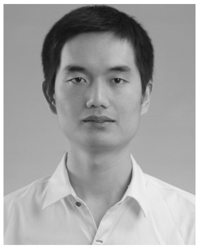
[57] H. Zhang, A. C. Berg, M. Maire, and J. Malik, "SVM-KNN: Discriminative nearest neighbor classification for visual category recognition," in *Proc. CVPR*, 2006, pp. 2126–2136.

[58] J. Donahue *et al.*, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 647–655.

[59] J. Jiang, R. Hu, Z. Wang, Z. Han, and J. Ma, "Facial image hallucination through coupled-layer neighbor embedding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 9, pp. 1674–1684, Sep. 2016, doi: 10.1109/TCSVT.2015.2433538.

Chunjie Zhang received the B.E. degree from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2006 and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2011.

He was an Engineer with the Henan Electric Power Research Institute, Zhengzhou, China, from 2011 to 2012. He held a post-doctoral position at the School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing, where he is currently an Assistant Professor with the School of Computer and Control Engineering. His research interests include image processing, machine learning, pattern recognition, and computer vision.



Guibo Zhu received the B.E. degree from Wuhan University, Wuhan, China, in 2009 and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2016.

He is an Assistant Professor with the Research Center for Brain-inspired Intelligence, Institute of Automation, Chinese Academy of Sciences. His research interests include pattern recognition and machine learning, computer vision, and brain-inspired intelligence in vision.

Chao Liang received the B.S. degree in automation from Huazhong University of Science and Technology, Wuhan, China, in 2006 and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2012.

He is an Assistant Professor with the National Engineering Research Center for Multimedia Software, School of Computer, Wuhan University, Wuhan. His research interests include multimedia content analysis, machine learning, computer vision, and pattern recognition.

Yifan Zhang received the B.E. degree in automation from Southeast University in 2004 and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2010.

He joined the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, where he is currently an Associate Professor. From 2011 to 2012, he was a Post-Doctoral Research Fellow with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY, USA. His research interests include probabilistic graphical models, activity recognition, and video semantic analysis.

Qingming Huang received the Ph.D. degree in computer science from Harbin Institute of Technology, Harbin, China, in 1994.

He was a Post-Doctoral Fellow with National University of Singapore, Singapore, from 1995 to 1996, and was with the Institute for Infocomm Research, Singapore, as a Research Staff Member from 1996 to 2002. He joined the Chinese Academy of Sciences, Beijing, China, in 2003 and is currently a Professor with the University of Chinese Academy of Sciences, Beijing. His research interests include image and video analysis, video coding, pattern recognition, and computer vision.

Qi Tian received the B.E. degree from Tsinghua University, Beijing, China, in 1992; the M.S. degree from Drexel University, Philadelphia, PA, USA, in 1996; and the Ph.D. degree in electrical and computer engineering from University of Illinois at Urbana–Champaign, Champaign, IL, USA, in 2002.

He is a Professor with the Department of Computer Science, University of Texas at San Antonio, San Antonio, TX, USA, and an Adjunct Professor with Zhejiang University, Hangzhou, China, and Xidian University, Xi'an, China. His research interests include multimedia information retrieval, computational systems biology, biometrics, and computer vision.